

TP 2 : Simulation de lois de probabilités

Introduction

Dans le logiciel R, on peut simuler très simplement à peu près toutes les lois classiques. Dans ce TP, on se propose, pour chaque loi usuelle, de simuler des échantillons, et de calculer la moyenne et la variance de ces échantillons, ainsi que faire des graphiques représentant ces lois. Il y a quatre commandes à connaître pour chaque loi :

- `rmaoi` : permet de simuler selon maloi
- `dmaoi` : permet de calculer la densité de maloi
- `pmaoi` : permet de calculer la fonction de répartition de maloi
- `qmaoi` : permet de calculer le quantile de maloi

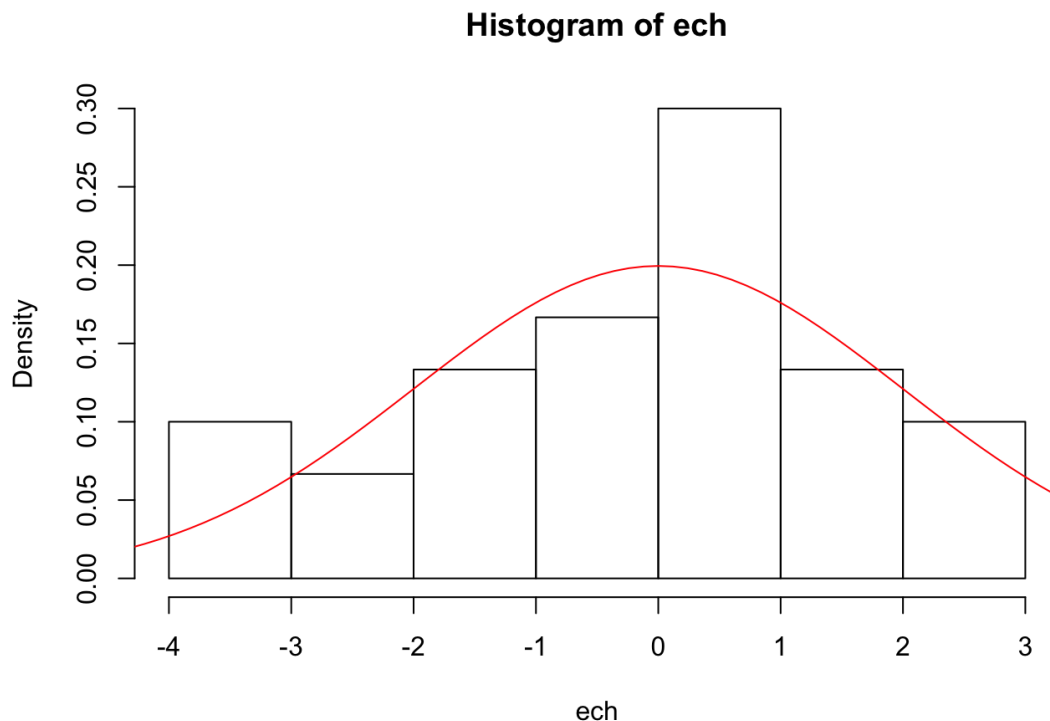
Exemple

Pour la loi normale :

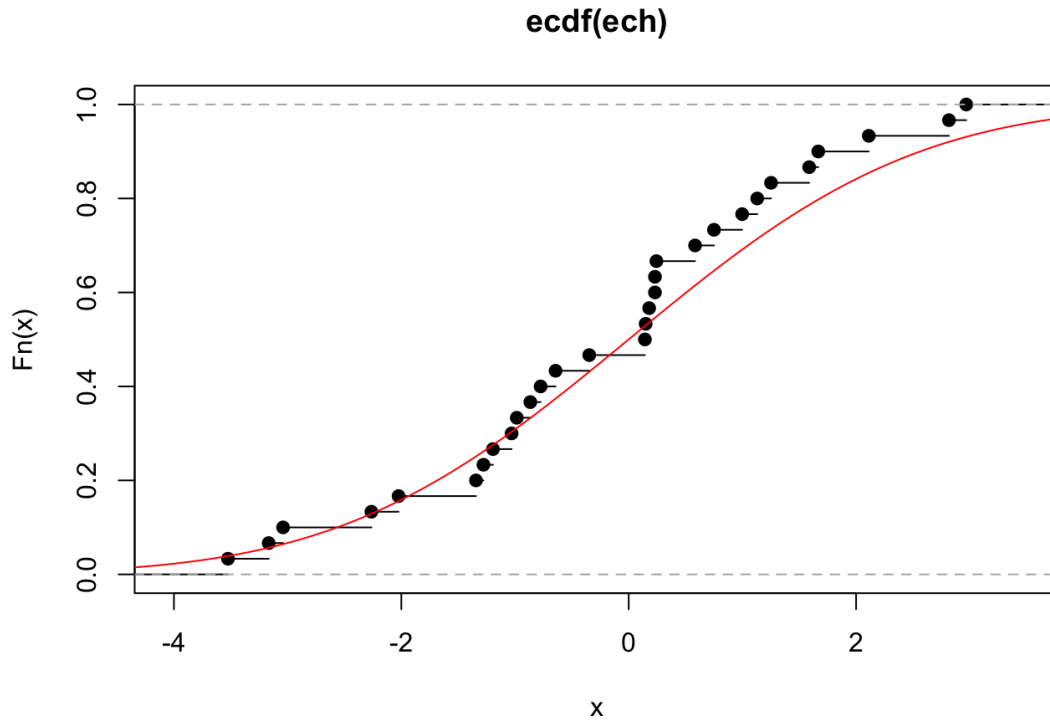
```
n<-30
ech<-rnorm(n,mean = 0,sd = 2) # simulation d'un échantillon d'une loi normale d'espérance 1 et de variance 4 (écart-type = 2)
head(ech,n = 20)
```

```
## [1]  0.2307228  1.1304828  1.6670783 -1.0301156 -3.1665735 -0.3471147
## [7]  2.1114359 -0.9845209  1.5871772 -1.2777375  0.7498350  0.2307736
## [13] 0.2435407  2.9686721 -2.0240850 -3.0401370 -1.3435749  0.1488665
## [19] -0.7749786 -1.1938214
```

```
hist(ech, freq=F)
curve(dnorm(x,mean = 0,sd = 2),from = -6,to = 6,ylab="densité",add=T,col="red")
```



```
plot(ecdf(ech))
curve(pnorm(x,mean = 0,sd = 2),from = -6,to = 6,ylab="Fonction de répartition",add=T,col="red")
```



```
quantile.normale<-qnorm(p = c(0.05,0.1,0.25,0.5,0.75,0.9,0.95),mean = 0,sd = 2)
quantile.echantillon<-quantile(x = ech,probs = c(0.05,0.1,0.25,0.5,0.75,0.9,0.95))
(rbind(quantile.echantillon,quantile.normale))
```

```
##              5%      10%      25%      50%      75%
## quantile.echantillon -3.109677 -2.340823 -1.152895 0.1456467 0.9357005
## quantile.normale     -3.289707 -2.563103 -1.348980 0.0000000 1.3489795
##              90%      95%
## quantile.echantillon 1.711514 2.499261
## quantile.normale     2.563103 3.289707
```

Lois discrètes

Loi uniforme discrète

Soit X une variable aléatoire distribuées selon une loi uniforme discrète sur $\{1, \dots, n\}$.

La distribution est définie comme suit :

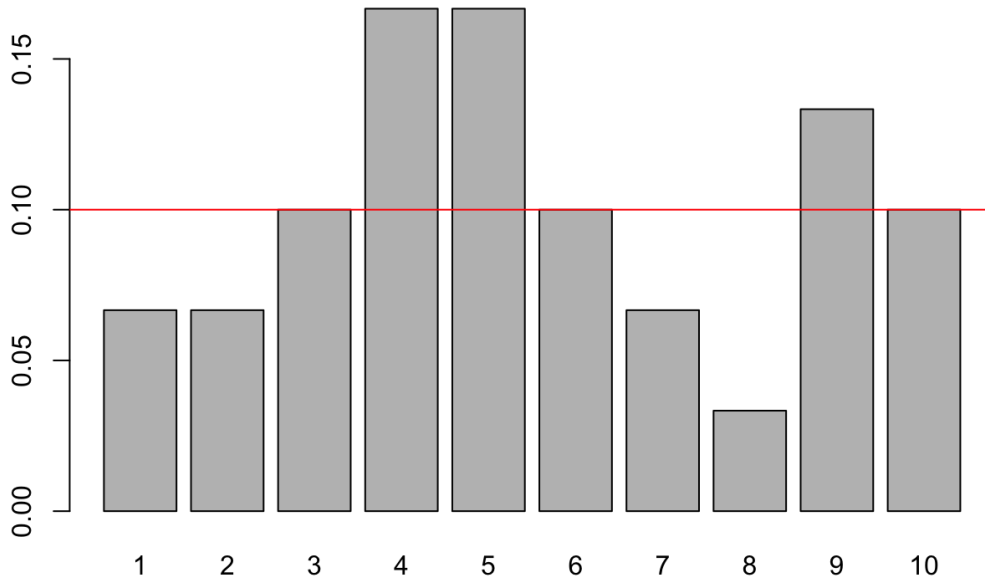
$$\forall k \in \{1, \dots, n\}, p(X = k) = \frac{1}{n}$$

En R,

```
n<-30
ech<-sample(1:10,n,replace=T)
table(ech)/n
```

```
## ech
##      1      2      3      4      5      6
## 0.06666667 0.06666667 0.10000000 0.16666667 0.16666667 0.10000000
##      7      8      9     10
## 0.06666667 0.03333333 0.13333333 0.10000000
```

```
barplot(table(ech)/n)
abline(h=1/10,col="red")
```



Loi de Bernoulli

Cette loi est notée $\mathcal{B}(p)$. On dit que $X \rightsquigarrow \mathcal{B}(p)$ si et seulement si X est une variable binaire telle que (on parle de loi succès/échec) :

$$P(X = 1) = p$$

et

$$P(X = 0) = 1 - p = q$$

C'est un cas particulier de la loi binomiale avec $n = 1$.

On peut vérifier facilement par le calcul que $\mathbb{E}[X] = p$ et $\mathbb{V}[X] = pq = p(1 - p)$.

```
n<- 30 # taille de l'échantillon
p<-0.3 # paramètre de la loi de Bernoulli
esp<-p
var<-p*(1-p)
ech<-rbinom(n,size = 1,prob = p) # size = 1 correspond à la loi de Bernoulli
table(ech)/n
```

```
## ech
## 0 1
## 0.6 0.4
```

```
paste("La moyenne de l'échantillon est ",mean(ech)," , sachant que l'espérance vaut",esp)
```

```
## [1] "La moyenne de l'échantillon est 0.4 , sachant que l'espérance vaut 0.3"
```

```
paste("La variance de l'échantillon est ",round(var(ech),2)," , sachant que la variance de la loi vaut",
p*(1-p))
```

```
## [1] "La variance de l'échantillon est 0.25 , sachant que la variance de la loi vaut 0.21"
```

Loi binomiale

On dit que X suit une loi binomiale, notée $\mathcal{B}(n, p)$.

On dit que cette loi est la somme de lois de Bernoulli indépendantes et de même paramètre, ce qui revient à compter le nombre de succès (nombre de 1) parmi les n expériences. Cela revient donc à faire n fois une expérience (qui a une probabilité p de réussir) identique, de façon indépendante. X représente donc le nombre de succès.

Pour résumer, soit $Y_i \rightsquigarrow \mathcal{B}(p)$ pour $i = 1, \dots, n$, où les Y_i sont indépendants. On note

$$X = \sum_{i=1}^n Y_i \rightsquigarrow \mathcal{B}(n, p)$$

Les valeurs possibles pour la variable X sont $\{0, \dots, n\}$ et

$$\forall k \in \{0, \dots, n\}, p_k = p(X = k) = C_n^k p^k (1-p)^{n-k}$$

On peut remarquer que la somme des probabilités est bien égale à 1 grâce à la formule du binôme de Newton :

$$(a + b)^n = \sum_{k=0}^n C_n^k a^k b^{n-k}$$

En prenant $a = p$ et $b = q = 1 - p$ on trouve :

$$\sum_{k=0}^n p_k = \sum_{k=0}^n C_n^k p^k (1-p)^{n-k} = (p + 1 - p)^n = 1^n = 1$$

L'espérance de cette loi est (grâce à la **linéarité de l'espérance**)

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[Y_i] = \sum_{i=1}^n p = np$$

La variance de cette loi est

$$\mathbb{V}[X] = \sum_{i=1}^n \mathbb{V}[Y_i] = \sum_{i=1}^n p(1-p) = np(1-p)$$

Attention, dans cette dernière formule, il ne faut pas croire que nous utilisons la linéarité de la variance ! **La variance n'est pas linéaire.** Pour rappel :

$$\mathbb{V}[aX] = a^2 \mathbb{V}[X]$$

et

$$\mathbb{V}[X_1 + X_2] = \mathbb{V}[X_1] + \mathbb{V}[X_2] + 2 * \text{cov}(X_1, X_2)$$

Enfin, si X_1 et X_2 sont indépendantes, alors

$$\text{cov}(X_1, X_2) = 0$$

En R, cela donne

```
N<- 100 # taille de l'échantillon
n<- 10L # premier paramètre de la loi
p<- 0.8 # deuxième paramètre de la loi

ech<-rbinom(N,n,p)
table(ech)/N
```

```
## ech
##   3   5   6   7   8   9  10
## 0.01 0.04 0.01 0.21 0.41 0.26 0.06
```

```
dbinom(x=0:n,size = n,prob = p)
```

```
## [1] 0.0000001024 0.0000040960 0.0000737280 0.0007864320 0.0055050240
## [6] 0.0264241152 0.0880803840 0.2013265920 0.3019898880 0.2684354560
## [11] 0.1073741824
```

Faites varier les différents paramètres de la loi binomiale et étudier ses propriétés.

Loi Géométrique

On dit que X suit une loi géométrique, notée $\mathcal{G}(p)$ avec $0 < p < 1$.

Cette loi correspond au schéma suivant. On répète, de façon indépendante, une même expérience (qui a une probabilité p de réussite et une probabilité $q = 1 - p$ d'échec) et on note X le rang du premier succès.

Les valeurs prises par X sont donc $X(\Omega) = \mathbb{N}^*$ et les probabilités associées sont :

$$\forall k \in \mathbb{N}^*, p_k = P(X = k) = (1 - p)^{k-1} p = q^{k-1} p$$

On remarque que ces probabilités p_k définissent une suite géométrique de raison q .

Pour rappel, si $0 < q < 1$ alors

$$\sum_{k=n_0}^n q^k = q^{n_0} \frac{1 - q^{n-n_0+1}}{1 - q}$$

et donc par passage à la limite

$$\sum_{k=n_0}^{+\infty} q^k = \frac{q^{n_0}}{1 - q}$$

En appliquant cette formule pour calculer la somme des probabilités, on trouve

$$\sum_{k=1}^{+\infty} p_k = \sum_{k=1}^{+\infty} q^{k-1} p = p \sum_{k=0}^{+\infty} q^k = p/(1 - q) = 1$$

L'espérance de cette loi est :

$$\mathbb{E}[X] = \sum_{k=1}^{+\infty} k p_k = p \sum_{k=1}^{+\infty} k q^{k-1} = p/(1 - q)^2 = 1/p$$

Le moment d'ordre 2 de cette loi est :

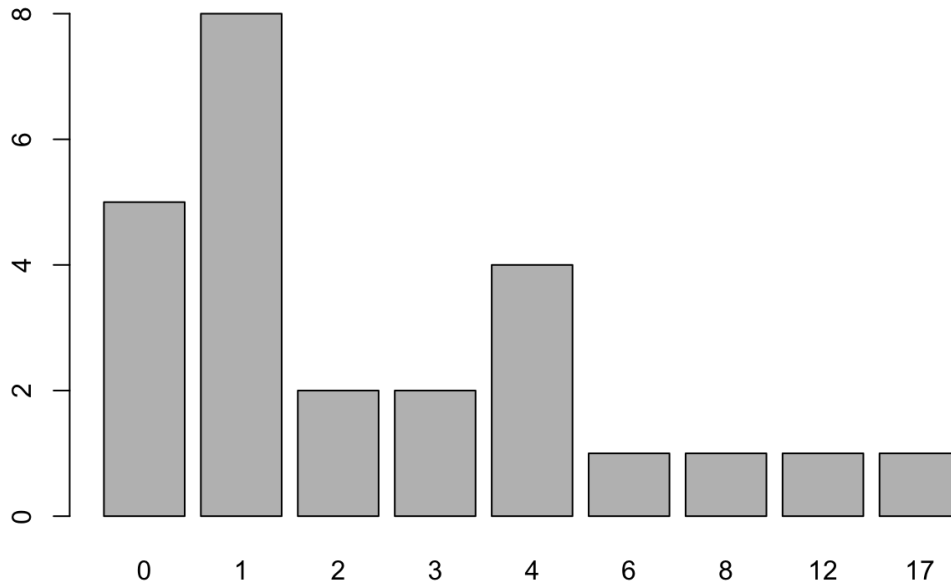
$$\mathbb{E}[X^2] = \sum_{k=1}^{+\infty} k^2 p_k = p \sum_{k=1}^{+\infty} k(k - 1 + 1) q^{k-1} = pq \sum_{k=1}^{+\infty} k(k - 1) q^{k-2} + p \sum_{k=1}^{+\infty} k q^{k-1} = pq \frac{2}{(1 - q)^3} + 1/p = \frac{2(1 - p) + p}{p^2} = \frac{1 + q}{p^2}$$

La variance vaut

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1 + q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}$$

En R, il faut faire attention car il y a un décalage d'une unité. La définition dans R pour une loi géométrique est le nombre d'échecs avant la première réussite. Ainsi, les valeurs prises pour cette variable sont dans \mathbb{N} et non pas \mathbb{N}^* .

```
n <- 25 # taille de l'échantillon
p <- 0.3 # paramètre de la loi
ech<-rgeom(n,p) # simulation de la loi
barplot(table(ech))
```



Redéfinir une nouvelle variable ou une nouvelle fonction qui correspond à notre paramétrisation de la loi géométrique

Loi de Poisson

On dit que X suit une loi de Poisson, notée $\mathcal{P}(\lambda)$ avec $\lambda > 0$.

On peut voir cette loi comme la loi des événements rares. Cette loi est une approximation de la loi binomiale, $\mathcal{B}(n, p)$ lorsque n est grand et p petit (i.e. pour des événements rares). On pose alors $\lambda = np$. En pratique, on dit que cette approximation est bonne dès que $n > 20$ et $p \leq 0.1$ et $np \leq 5$.

A l'aide de simulation, vérifier que la loi binomiale peut être approchée par la loi de poisson.

Les valeurs possibles pour X sont dans \mathbb{N} et les probabilités associées :

$$\forall k \in \mathbb{N}, p_k = P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

On pourra vérifier (grâce à la définition de l'exponentielle $e^x = \sum_{k=0}^{+\infty} \frac{x^k}{k!}$) que

$$\sum_{k=0}^{+\infty} p_k = 1$$

L'espérance d'une loi de Poisson de paramètre λ est :

$$\mathbb{E}[X] = \sum_{k=0}^{+\infty} k p_k = \sum_{k=1}^{+\infty} k p_k = \lambda e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda$$

Le moment d'ordre 2 vaut

$$\mathbb{E}[X^2] = \sum_{k=0}^{+\infty} k^2 p_k = \sum_{k=1}^{+\infty} k^2 p_k = \lambda e^{-\lambda} \sum_{k=1}^{+\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} = \lambda^2 e^{-\lambda} \sum_{k=2}^{+\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda^2 + \lambda$$

et donc la variance est égale à

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

En R, cela donne

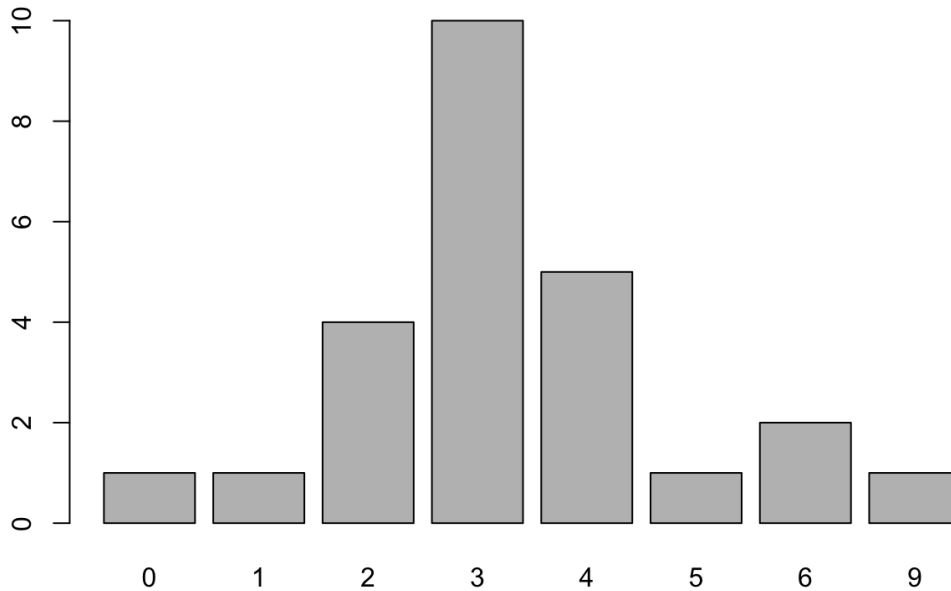
```
lambda <- 3 # paramètre de la loi de Poisson
n <- 25 # taille de l'échantillon
ech <- rpois(n,lambda) # simulation de la loi de Poisson
mean(ech)
```

```
## [1] 3.4
```

```
var(ech)
```

```
## [1] 3.166667
```

```
barplot(table(ech))
```



Simulez différentes lois de Poisson en faisant varier λ ainsi que la taille de l'échantillon.

Pour toutes ces lois, comparer la fonction de répartition empirique avec la théorique.

Lois continues

Introduction

Lisez la partie du cours relative aux lois continues.

Loi uniforme sur $[0, 1]$, notée $U_{[0,1]}$

On dit que X suit une loi uniforme sur $[0, 1]$ ssi sa fonction densité est :

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} 1 & \text{si } x \in [0, 1] \\ 0 & \text{sinon} \end{cases}$$

La fonction de répartition d'une loi uniforme sur $[0, 1]$ est

$$\forall x \in \mathbb{R}, F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \in [0, 1] \\ 1 & \text{si } x > 1 \end{cases}$$

L'espérance de la loi uniforme $U_{[0,1]}$ est

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x) dx = \int_0^1 x dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

Le moment d'ordre 2 est

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \int_0^1 x^2 dx = \left[\frac{x^3}{3} \right]_0^1 = \frac{1}{3},$$

et donc la variance vaut

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1/3 - 1/4 = 1/12.$$

```
n <- 30
ech <- runif(n)
mean(ech)
```

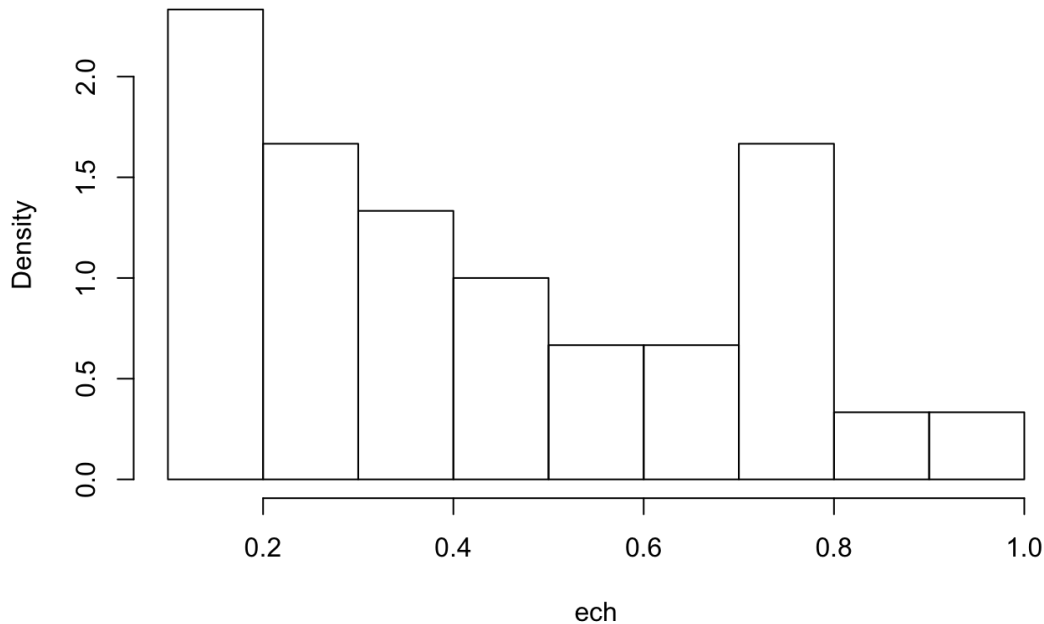
```
## [1] 0.4322295
```

```
var(ech)
```

```
## [1] 0.05778539
```

```
hist(ech, freq=F)
```

Histogram of ech



Loi uniforme sur $[a, b]$, notée $U_{[a,b]}$, avec $(a < b)$

On dit que $X \rightsquigarrow U_{[a,b]}$ ssi sa fonction densité est

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{sinon} \end{cases}$$

La fonction de répartition d'une loi uniforme sur $[a, b]$ est

$$\forall x \in \mathbb{R}, F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } x \in [a, b] \\ 1 & \text{si } x > b \end{cases}$$

L'espérance vaut

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

Le moment d'ordre 2 est

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \int_a^b \frac{x^2}{b-a} dx = \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3},$$

et donc la variance vaut

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{a^2 - 2ab + b^2}{12} = \frac{(b-a)^2}{12}$$

Théorème : Si $X \rightsquigarrow U_{[0,1]}$ alors $Y = (b-a)X + a \rightsquigarrow U_{[a,b]}$.

En R, il n'est pas utile de connaître ce théorème car la fonction *runif* permet de passer les bornes a et b en arguments.

```
n<-30
ech<-runif(n,2,5) # simulation d'un échantillon d'une loi uniforme sur [2,5]
mean(ech)
```

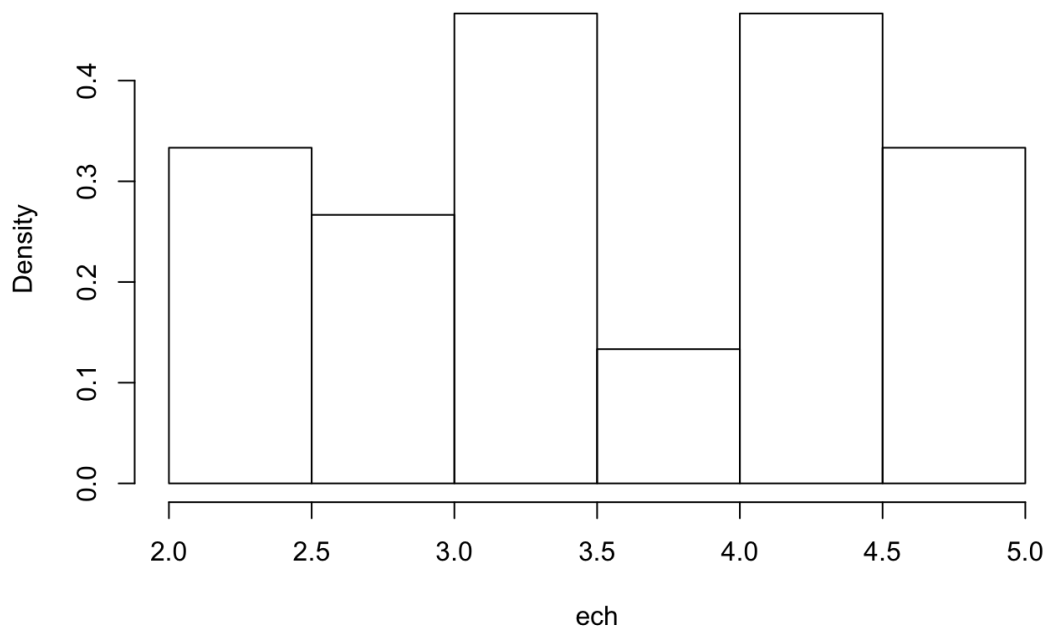
```
## [1] 3.495684
```

```
var(ech)
```

```
## [1] 0.7500752
```

```
hist(ech,freq=F)
```

Histogram of ech



Loi Normale centrée réduite, notée $\mathcal{N}(0, 1)$

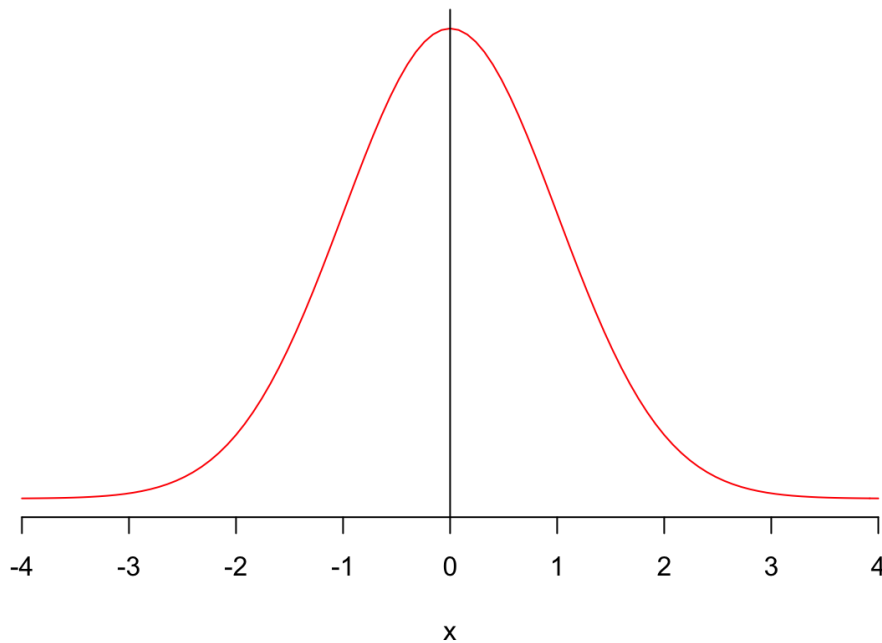
Si $X \rightsquigarrow \mathcal{N}(0, 1)$ alors la fonction densité de X est

$$\forall x \in \mathbb{R}, f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Le premier paramètre de la loi normale est l'espérance de la loi et le second correspond à la variance de la loi. On peut remarquer que cette fonction est paire, i.e. $\forall x \in \mathbb{R}, f(-x) = f(x)$, ce qui implique une symétrie par rapport à l'axe des ordonnées.

```
curve(dnorm, from = -4, to = 4, col="red", main="Densité de la loi Normale centrée réduite", ylab="", axes=F)
axis(1, xaxp=c(-4, 4, 8))
axis(2, pos=0, c(-1, 1))
```

Densité de la loi Normale centrée réduite



On peut également remarquer que la courbe s'approche très rapidement de l'axe des abscisses dès lors que l'on s'éloigne de l'espérance (ici 0).

```
n<-100
ech<-rnorm(n) # simulation d'une loi normale centrée réduite
mean(ech)
```

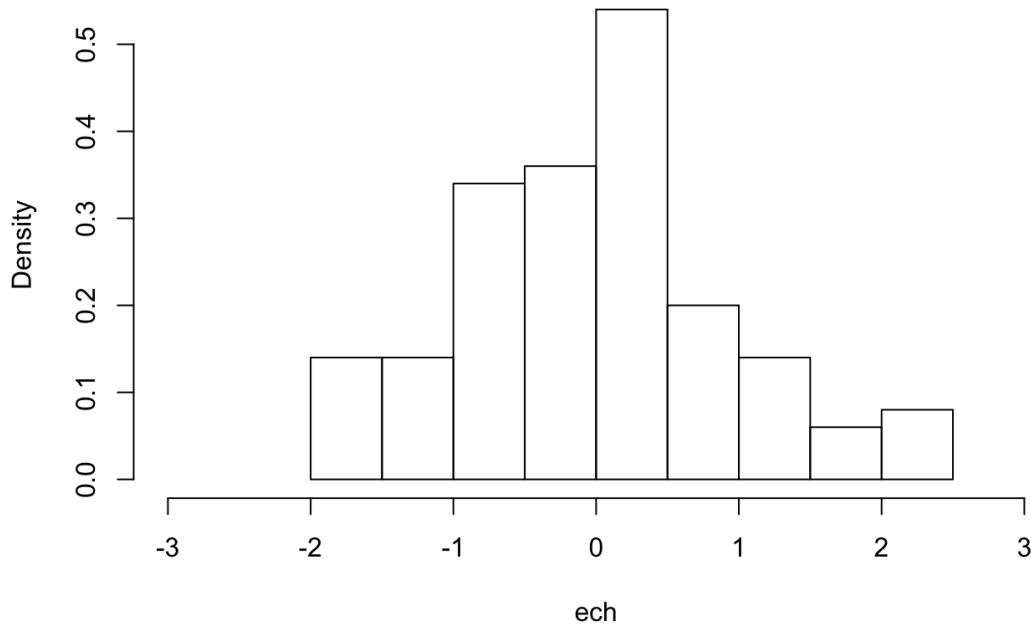
```
## [1] -0.02140691
```

```
var(ech)
```

```
## [1] 0.9093207
```

```
hist(ech, freq=F, breaks=10, xlim = c(-3, 3))
```

Histogram of ech



Vérifier de plusieurs manières que

$$\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}.$$

L'espérance de la loi normale centrée réduite vaut

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} xe^{-\frac{x^2}{2}} dx = 0$$

Le moment d'ordre 2 (qui est aussi la variance car l'espérance est nulle) se calcule par une intégration par parties

$$\mathbb{V}[X] = \mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} dx = \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} x^2 e^{-\frac{x^2}{2}} dx$$

Or

$$\int_0^t x^2 e^{-\frac{x^2}{2}} dx = \left[-xe^{-\frac{x^2}{2}} \right]_0^t + \int_0^t e^{-\frac{x^2}{2}} dx \xrightarrow{t \rightarrow +\infty} \int_0^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{\frac{\pi}{2}}$$

et donc

$$\mathbb{V}[X] = 1$$

Loi Normale $\mathcal{N}(\mu, \sigma^2)$

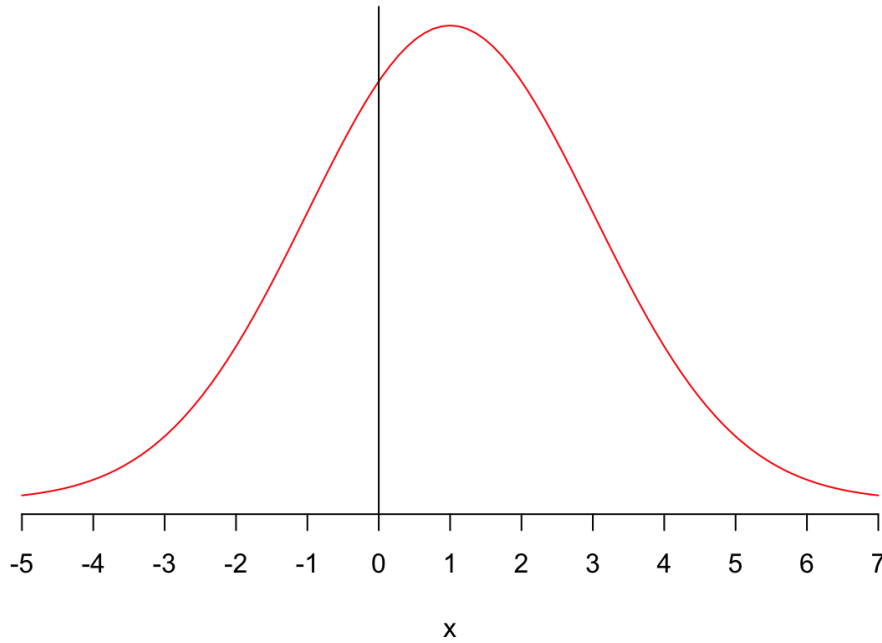
On dit que $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ si sa fonction densité est :

$$\forall x \in \mathbb{R}, f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

On pourra vérifier que l'intégrale de cette fonction sur \mathbb{R} vaut bien 1. μ représente l'espérance de la loi et σ^2 la variance.

```
curve(dnorm(x,mean=1,sd=2),from = -5,to = 7,col="red",main="Densité de la loi N(1,4)",axes=F,ylab="") #
Attention, on rentre l'écart-type et non la variance !!!
axis(1,xaxp = c(-5,7,12))
axis(2,pos=0,c(-1,1))
```

Densité de la loi $N(1,4)$



Que remarquez-vous ?

Théorème : Si $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ alors $\frac{X - \mu}{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$

Ce théorème est extrêmement important car il dit qu'on est toujours capable de se ramener à une loi normale centrée réduite (encore faut-il connaître l'espérance et la variance de la loi).

Mettre en oeuvre un programme permettant de vérifier ce théorème par la simulation.

Exercices sur la loi normale

Distribution des valeurs

Pour simuler n valeurs aléatoirement suivant la loi normale de moyenne m et d'écart-type s on utilise sous R la commande

```
rnorm(20,0,1) # simulation d'un échantillon de taille 20 d'une loi N(0,1)
```

```
## [1] -1.7032087 -2.0534813 -0.4378787  2.5270446 -0.5457850  0.9654907
## [7] -0.4282005  1.1025392 -1.1312043  0.2999487  0.8402349 -1.0814214
## [13]  0.5904264 -0.9485972 -0.4725433 -0.3680458 -1.1901503  0.1842897
## [19]  0.3722096 -1.0157028
```

```
rnorm(20,1,1) # simulation d'un échantillon de taille 20 d'une loi N(1,1)
```

```
## [1]  1.8867801  1.6159368  0.6998990  1.7955722  1.1335467  1.5035369
## [7]  0.3348979  1.7247437  1.2954924 -0.0586502  0.2012321  1.8088699
## [13]  0.1935927  0.3701374  1.3362243  2.5321147 -0.1263768  1.1520324
## [19]  1.7443802 -0.7022098
```

```
rnorm(20,2,1) # simulation d'un échantillon de taille 20 d'une loi N(2,1)
```

```
## [1]  1.1594811  3.0008013  0.5639313  2.9202225  3.9640892  3.1307478  4.1760582
## [8]  1.5139560  1.6798439  3.6444234  2.5063018  2.8740404  1.7363008  2.0191012
## [15]  2.8671983  1.5484510  1.1397431  3.0331141  1.1176564  0.7842981
```

```
rnorm(20,0,2) # simulation d'un échantillon de taille 20 d'une loi N(0,1)
```

```
## [1] -3.9318519 -0.7959969 -2.6530789 0.9173994 -3.0609998 3.8701073
## [7] -0.6678423 -0.6091488 0.7059081 2.9011270 2.6079106 -1.8114475
## [13] -1.4332422 -0.3484073 -1.2058270 -2.4926781 1.8304813 0.7455695
## [19] 0.8558886 1.1692576
```

```
rnorm(20,0,4)
```

```
## [1] 1.1113208 0.2002640 2.0290919 -5.9608327 -0.7393258 -2.3503499
## [7] -2.9548804 -1.0685149 1.5091304 3.0101102 -1.9021075 -4.5605041
## [13] 10.4379319 -1.3745206 2.8863127 -2.6154390 0.7345934 2.6945854
## [19] -1.5971151 1.9429857
```

Pour chacun des échantillons ci-dessus remplir le tableau suivant :

$n = 20$	$m = 0$ et $s = 1$	$m = 1$ et $s = 1$	$m = 2$ et $s = 1$	$m = 0$ et $s = 2$	$m = 0$ et $s = 4$
----------	--------------------	--------------------	--------------------	--------------------	--------------------

nb de valeurs < -2

nb de valeurs < 0

nb de valeurs = 0

nb de valeurs >2

Pour toute la suite du TP, voici un script R qui vous guidera pour répondre aux questions posées.

```
oldpar<-par() # mise en memoire des param?tres graphiques par d?faut
par(xaxs="i",yaxs="i") # re-d?finition des styles d'axes pour les graphiques de ce TP3

#####
# QUESTION 1
#####
# Echantillons de taille n=2000 individus avec diff?rents param?tres
n<-2000
nd1<-rnorm(n,0,1)
nd2<-rnorm(n,1,1)
nd3<-rnorm(n,2,1)
nd4<-rnorm(n,0,2)
nd5<-rnorm(n,0,4)

nd_en_col<-cbind(nd1,nd2,nd3,nd4,nd5) # les echantillons sont mis en ensemble, colonne par colonne
colSums(nd_en_col< -2)/n # pour chaque colonne on compte le nombre de valeurs qui est en dessous de -2 et on ramène cela à un pourcentage
```

```
## nd1 nd2 nd3 nd4 nd5
## 0.0225 0.0010 0.0000 0.1595 0.3070
```

Question 1 : avec les mêmes valeurs des moyennes et des écart-types, on simule maintenant des échantillons de taille $n = 2000$ individus.

Pour chacun de ces échantillons remplir le tableau suivant en utilisant les commandes R dans le script :

$n = 2000$	$m = 0$ et $s = 1$	$m = 1$ et $s = 1$	$m = 2$ et $s = 1$	$m = 0$ et $s = 2$	$m = 0$ et $s = 4$
------------	--------------------	--------------------	--------------------	--------------------	--------------------

nb de valeurs < -2

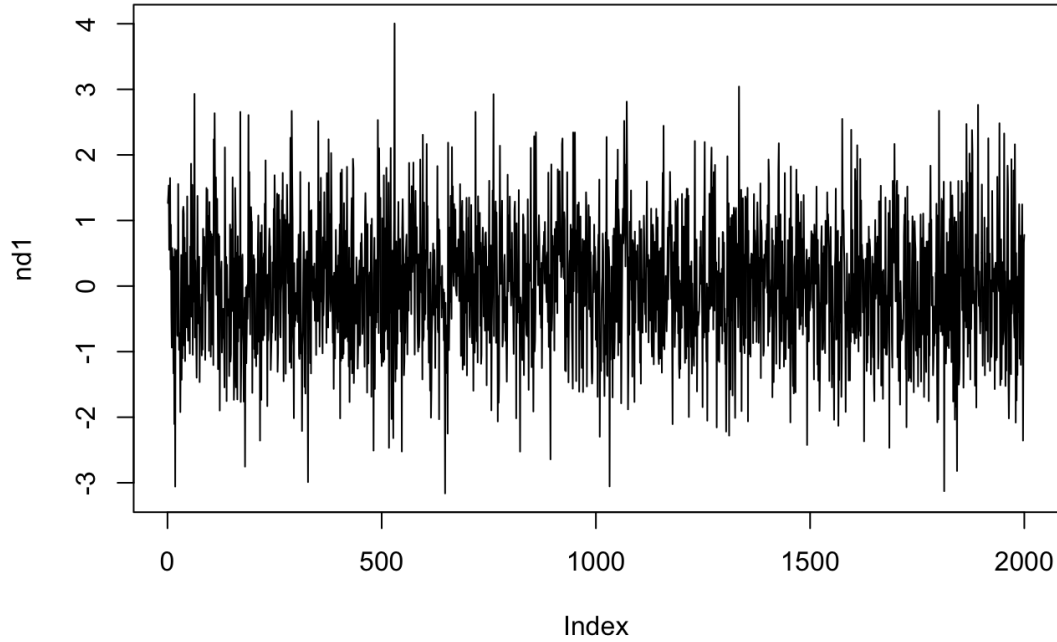
nb de valeurs < 0

nb de valeurs = 0

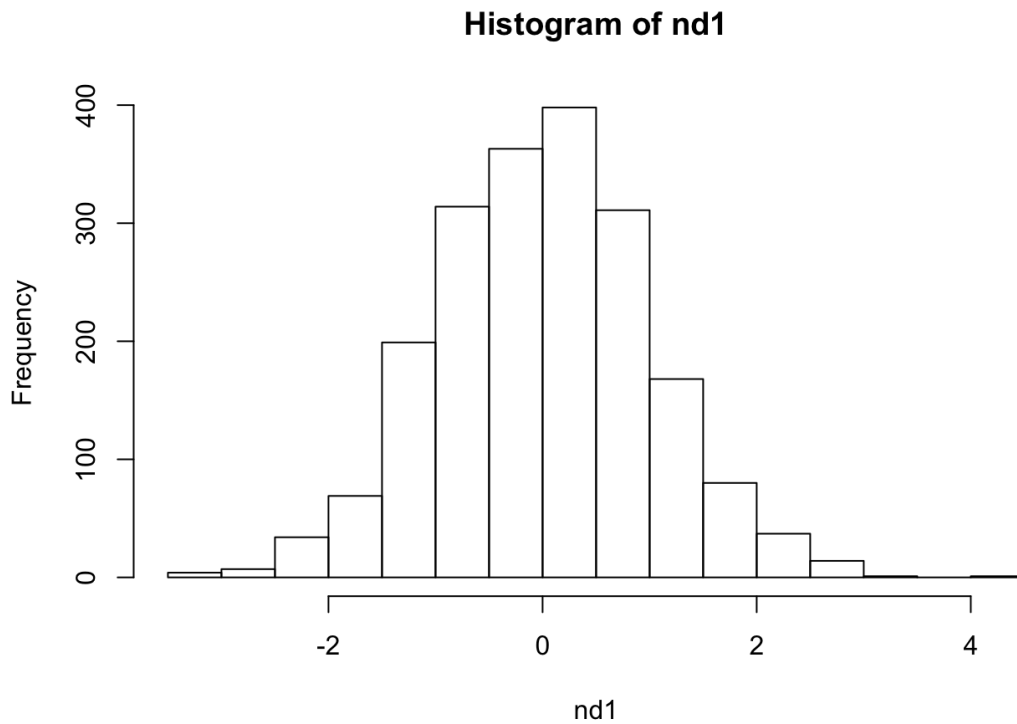
nb de valeurs >2

Question 2 : on considère uniquement le premier échantillon $nd1$. Que trace la fonction *plot* dans le script ? (axe des $x = ?$ et axe des $y = ?$) Pour dessiner la répartition des valeurs on trace un histogramme.

```
#####  
# QUESTION 2  
#####  
# Tracer les valeurs  
plot(nd1,type="l")
```

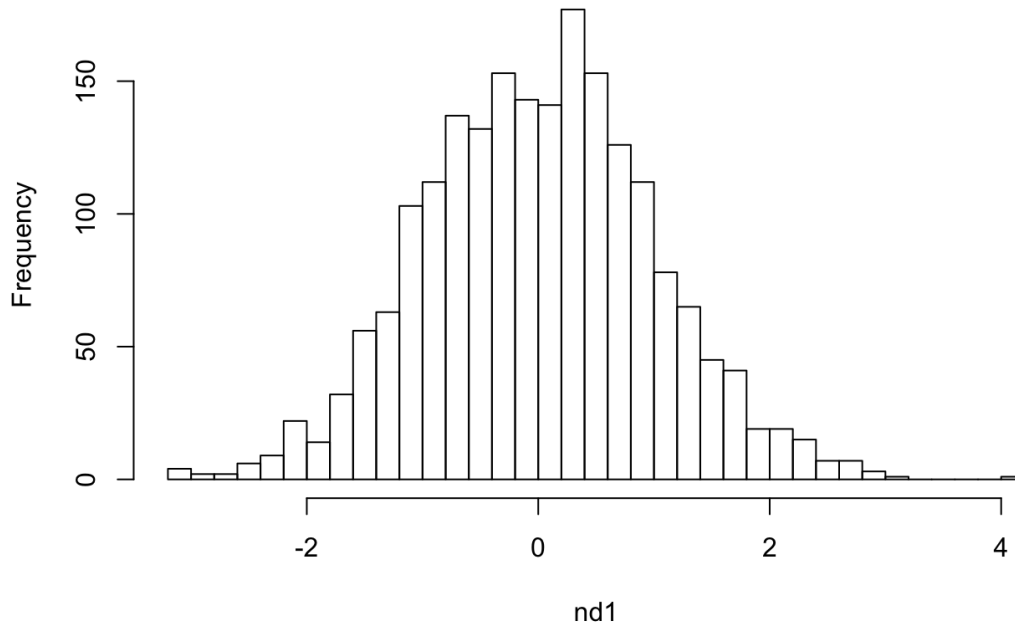


```
# Executer les commandes ci dessous et quelles sont les différences ?  
hist(nd1)
```



```
hist(nd1,freq=TRUE,breaks=40)  
hist(nd1,freq=TRUE,breaks=50)
```

Histogram of nd1



Question 3 : on veut comparer les histogrammes des 5 échantillons. Lancer les commandes de la question 3 qui permettent d'afficher les histogrammes et de superposer en rouge la densité théorique associée (pour mieux voir les graphiques cliquer sur zoom dans le menu *Plots*). Comparer les graphiques des échantillons *nd1*, *nd2* et *nd3* : qu'est-ce qui change ?

Comparer les graphiques des échantillons *nd1*, *nd4* et *nd5* : qu'est-ce qui change ?

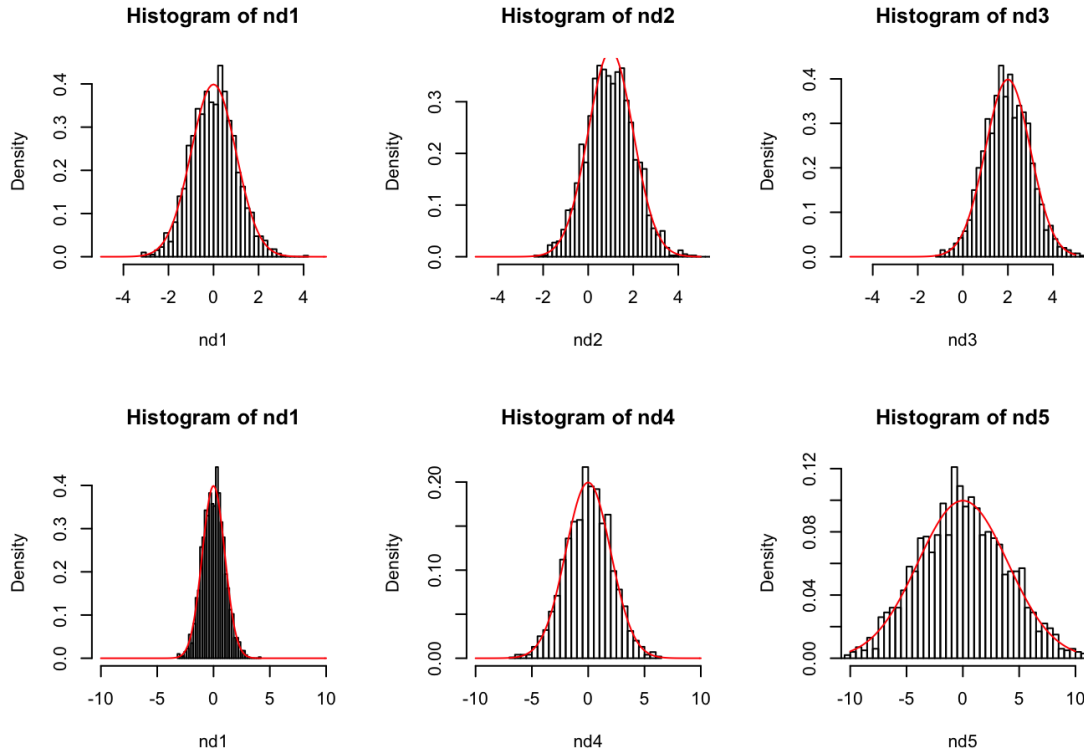
```
#####
# QUESTION 3
#####
# On trace plusieurs histogrammes pour comparer la répartition des valeurs
par(mfrow=c(2,3)) # 6 graphes dans une seule fenetre
                  # sur 2 lignes et 3 colonnes

infx<- -5        # valeur min sur l'axe des x
supx<- 5         # valeur max sur l'axe des x
x<-seq(infx,supx,0.1) # creation d'un vecteur du min au max par pas de 0.1

hist(nd1,freq=FALSE,breaks=40,xlim=c(infx,supx))
lines(x,dnorm(x,0,1),col="red") # on ajoute a l'histogramme des valeurs
                                # la courbe theorique de la densite

hist(nd2,freq=FALSE,breaks=40,xlim=c(infx,supx))
lines(x,dnorm(x,1,1),col="red")
hist(nd3,freq=FALSE,breaks=40,xlim=c(infx,supx))
lines(x,dnorm(x,2,1),col="red")

# seconde ligne de graphiques
infx<- -10
supx<- 10
x<-seq(infx,supx,0.1)
hist(nd1,freq=FALSE,breaks=40,xlim=c(infx,supx))
lines(x,dnorm(x,0,1),col="red")
hist(nd4,freq=FALSE,breaks=40,xlim=c(infx,supx))
lines(x,dnorm(x,0,2),col="red")
hist(nd5,freq=FALSE,breaks=40,xlim=c(infx,supx))
lines(x,dnorm(x,0,4),col="red")
```

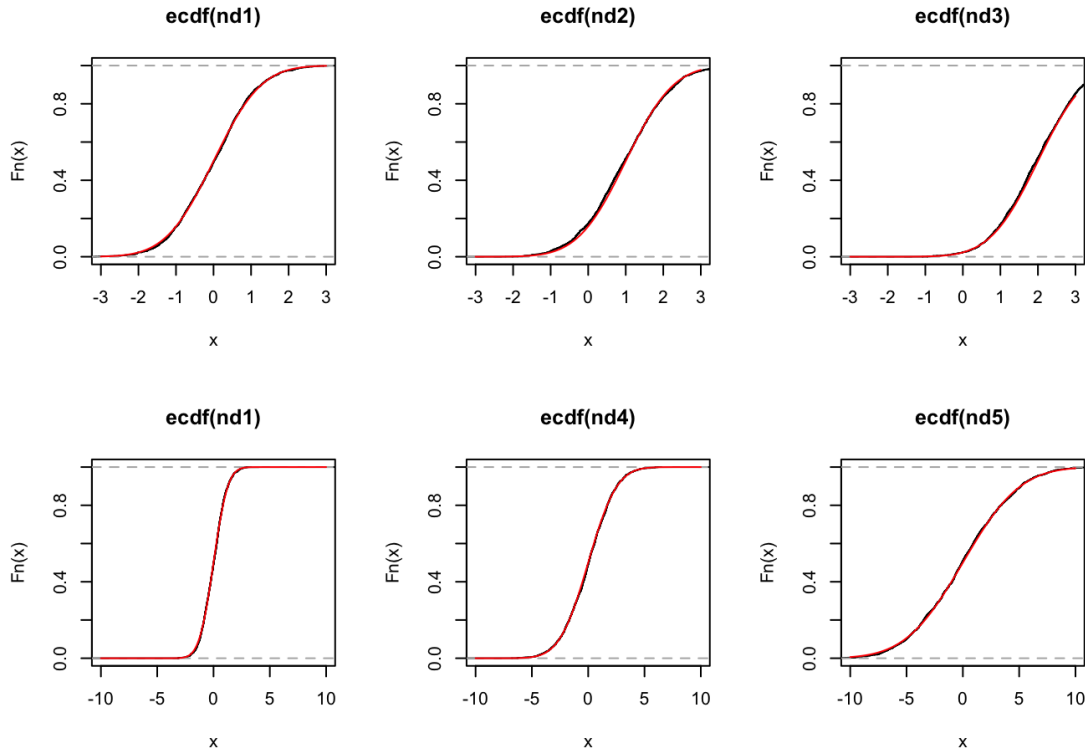


Question 4 : on veut maintenant comparer les fonctions de répartition des 5 échantillons. La fonction `plot(ecdf(x))` permet de tracer la fonction de répartition empirique associée aux valeurs du vecteur `x`. Lancer les commandes de la question 4; sur chaque fonction de répartition empirique (en noir) on trace la fonction de répartition théorique en rouge.

Comparer les graphiques des échantillons `nd1`, `nd2` et `nd3` : qu'est-ce qui change ?

Comparer les graphiques des échantillons `nd1`, `nd4` et `nd5` : qu'est-ce qui change ?

```
#####
# QUESTION 4
#####
# On trace les fonctions de répartition des différents
# échantillons pour comparer la répartition des valeurs
par(mfrow=c(2,3))
infx<- -3
supx<- 3
x<-seq(infx,supx,0.1)
plot(ecdf(nd1),xlim=c(infx,supx))
lines(x,pnorm(x,0,1),col="red")
plot(ecdf(nd2),xlim=c(infx,supx))
lines(x,pnorm(x,1,1),col="red")
plot(ecdf(nd3),xlim=c(infx,supx))
lines(x,pnorm(x,2,1),col="red")
infx<- -10
supx<- 10
x<-seq(infx,supx,0.1)
plot(ecdf(nd1),xlim=c(infx,supx))
lines(x,pnorm(x,0,1),col="red")
plot(ecdf(nd4),xlim=c(infx,supx))
lines(x,pnorm(x,0,2),col="red")
plot(ecdf(nd5),xlim=c(infx,supx))
lines(x,pnorm(x,0,4),col="red")
```

```
par(mfrow=c(1,1))
```

Question 5 : on veut montrer le lien entre la densité et la fonction de répartition.

Donner ci-dessous la densité de la loi normale de moyenne 0 et de variance 1 :

$$f(x) =$$

Quel est le lien entre la densité et la fonction de répartition ?

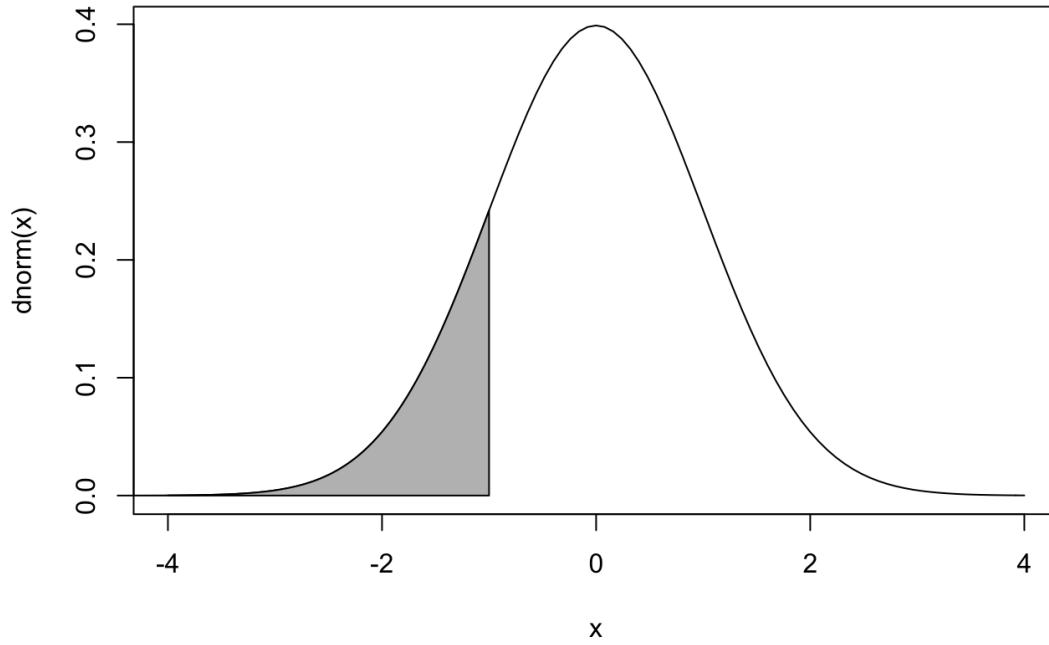
$$f(x) =$$

$$F(t) =$$

La valeur au point t de la fonction de répartition (théorique) $F(t)$ d'une loi gaussienne de moyenne m et d'écart-type s est donnée par la commande `pnorm(t,m,s)` de R.

Lancer le script de la question 5 et décrire ce qu'il permet d'observer.

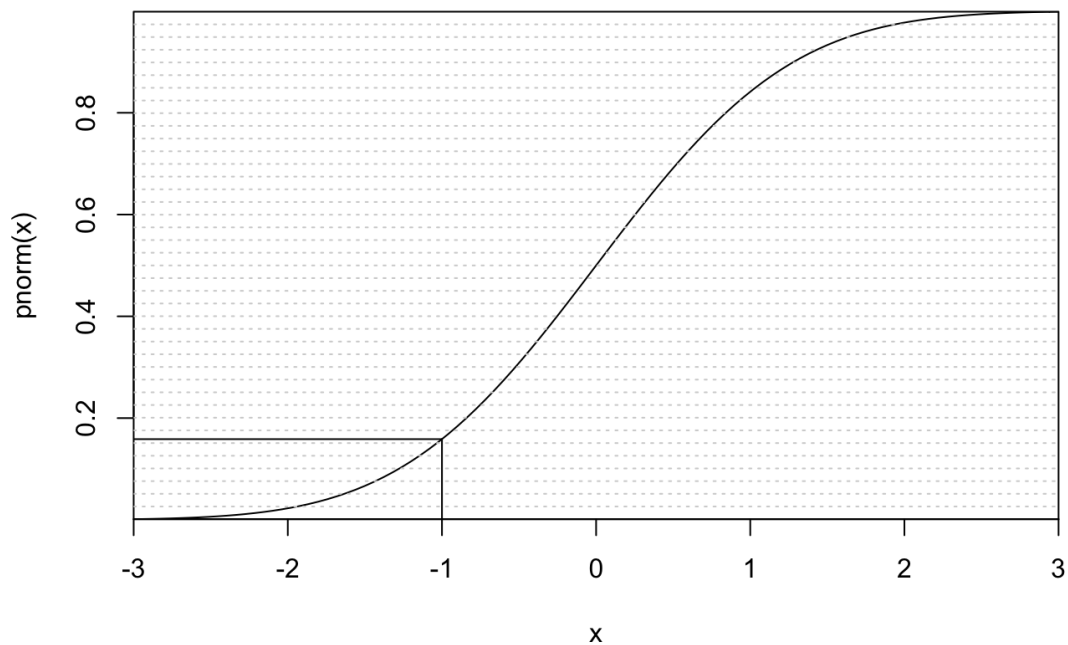
```
#####
# QUESTION 5
#####
curve(dnorm,from=-4,to=+4)
# creation d'une fonction pour colorier l'aire sous la courbe d'une
# fonction <fun> entre <xmin> et <xmax>
shade_under_curve <- function(fun, xmin, xmax, length=100){
  xvals <- seq(xmin, xmax, length=length)
  dvals <- match.fun(fun)(xvals)
  polygon(c(xvals,rev(xvals)),c(rep(0,length),rev(dvals)),col="gray")
}
shade_under_curve(dnorm,-5,-1)
```



```
##### valeur de l'aire sous la courbe
integrate(dnorm,-Inf,-1)
```

```
## 0.1586553 with absolute error < 4.8e-07
```

```
##### valeur de F(-1) à lire sur le graphe
curve(pnorm,from=-3,to=3,xaxs="i",yaxs="i")
grid(NA,40,lwd=1)
segments(-1,0,-1,pnorm(-1))
segments(-1,pnorm(-1),-3,pnorm(-1))
```



```
##### valeur de F(-1) par la fonction pnorm
pnorm(-1)
```

```
## [1] 0.1586553
```

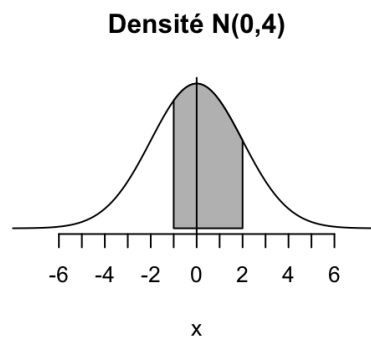
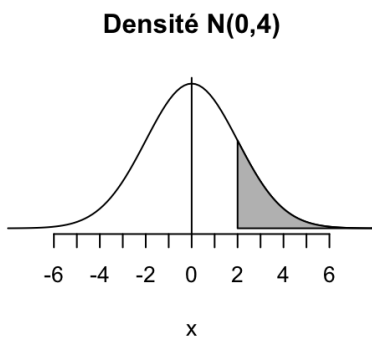
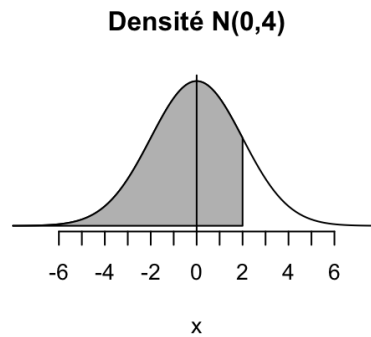
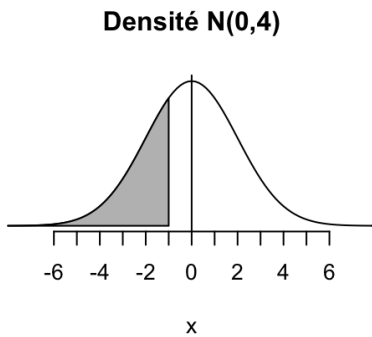
Question 6 : on a tracé ci dessous la courbe de densité d'une loi gaussienne centrée (de moyenne 0) et de variance 4 et l'on a hachuré différentes aires sous la courbe.

```
par(mfrow=c(2,2))
N04<-function(x) dnorm(x,0,2)
curve(N04,from = -8,to = 8,ylab="",axes=FALSE,main="Densité N(0,4)")
shade_under_curve(N04,-8,-1)
axis(1,xaxp=c(-6,6,12))
axis(2,pos=0,c(-1,1))

curve(N04,from = -8,to = 8,ylab="",axes=F,main="Densité N(0,4)")
shade_under_curve(N04,-8,2)
axis(1,xaxp=c(-6,6,12))
axis(2,pos=0,c(-1,1))

curve(N04,from = -8,to = 8,ylab="",axes=F,main="Densité N(0,4)")
shade_under_curve(N04,2,8)
axis(1,xaxp=c(-6,6,12))
axis(2,pos=0,c(-1,1))

curve(N04,from = -8,to = 8,ylab="",axes=F,main="Densité N(0,4)")
shade_under_curve(N04,-1,2)
axis(1,xaxp=c(-6,6,12))
axis(2,pos=0,c(-1,1))
```



Compléter le tableau ci-dessous :

Aire de la figure	Proba représentée par l'aire	Ecriture avec $F(t)$	Valeur numérique pnorm
1	$P(X < -1)$	$F(-1)$	
2			
3			
4			

Question 6 : soit X une variable aléatoire gaussienne de moyenne 15 et de variance 9.

Quelle est la probabilité que $X \leq 17$? (utiliser la fonction `pnorm`)

Quelle est la probabilité que $X > 7$? Enfin quelle est la probabilité que $7 < X < 15$?

Lien entre lois normales

Soit X une variable aléatoire distribuée suivant la loi normale d'espérance 0 et d'écart-type 1.

On souhaite montrer que les variables $X_1 = X + 1$, $X_2 = X - 1$, $X_3 = 2X$ et $X_4 = X/2$ suivent aussi une loi normale. Pour cela on simule un grand nombre de valeurs suivant la loi normale $N(0, 1)$ et l'on construit les nouvelles variables. Étudier la distribution de ces nouvelles variables.

```
x<-rnorm(10000,0,1)
x1<-x+1
x2<-x-1
x3<-x*2
x4<-x/2
```

Soit maintenant Y une variable aléatoire de moyenne 4 et de variance 3, quelle est la distribution de la variable $Z = \frac{Y-2}{\sqrt{3}}$?

Sommes de variables aléatoires gaussiennes

En vous aidant des résultats du script suivant, donner la loi de $X + Y$ et de $X - Y$ si $X \rightsquigarrow N(m_1, \sigma_1^2)$ et $Y \rightsquigarrow N(m_2, \sigma_2^2)$.

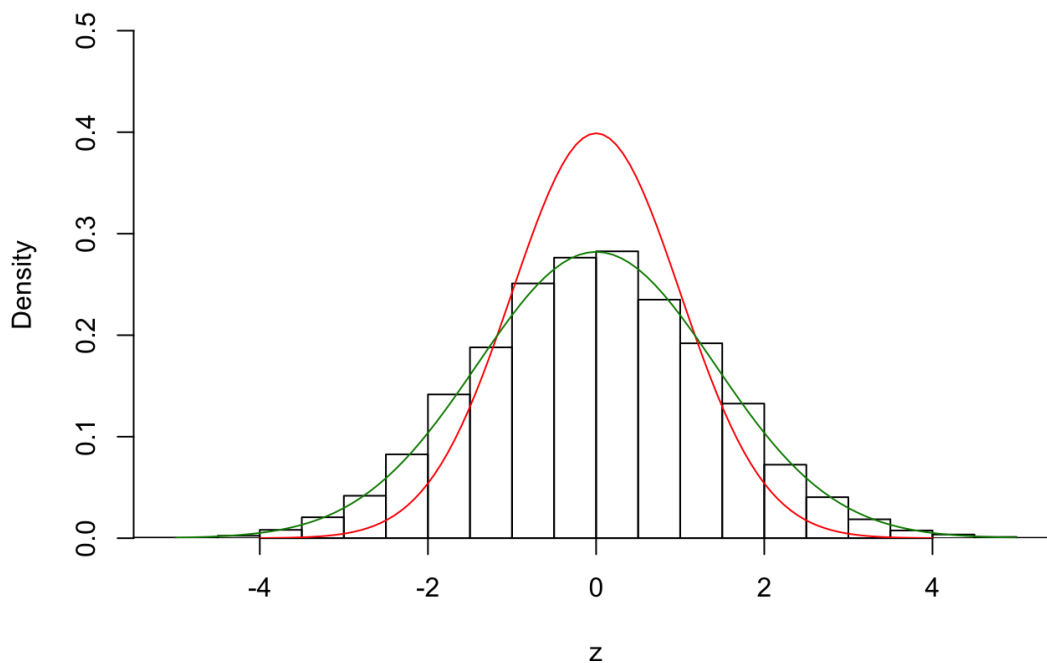
```
old_par<-par()
par(xaxs="i",yaxs="i")

# Deux variables gaussiennes de m?me loi
x<-rnorm(10000,0,1)
y<-rnorm(10000,0,1)

z<-x+y
hist(z,breaks=20,freq=FALSE,ylim=c(0,0.5))
curve(dnorm,from=-4,to=4,add=TRUE,col="red") # la loi N(0,1)

x<-seq(-5,5,0.1)
lines(x,dnorm(x,0,sqrt(2)),col="green4") #la loi propos?e
```

Histogram of z

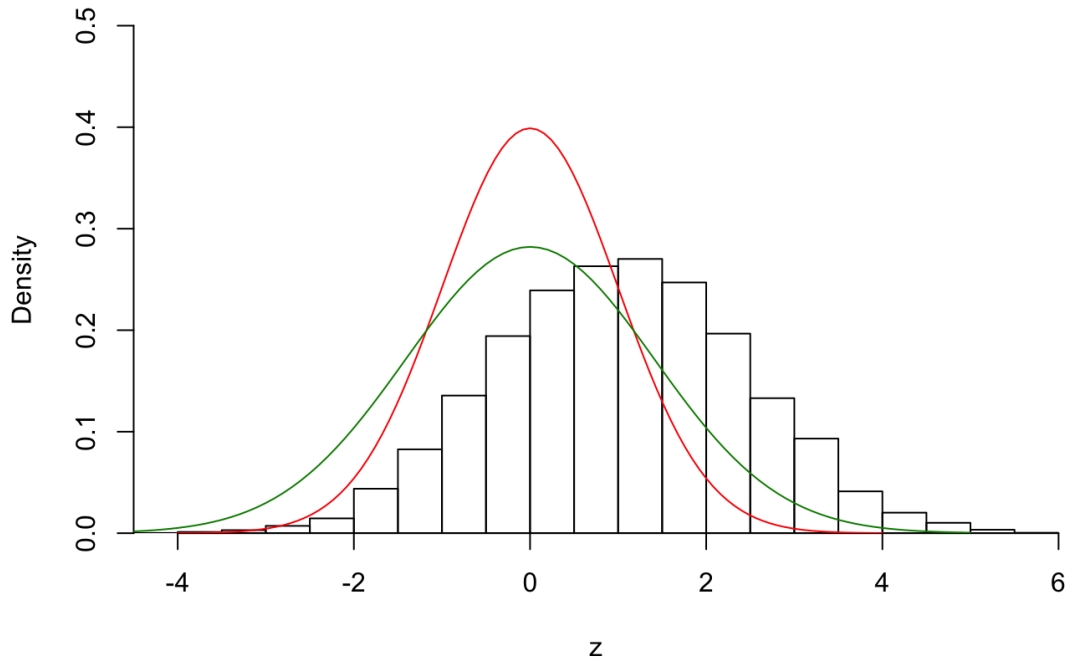


```
# Deux variables gaussiennes de meme ecart type mais pas de meme esperance
x<-rnorm(10000,0,1)
y<-rnorm(10000,1,1)

z<-x+y
hist(z,breaks=20,freq=FALSE,ylim=c(0,0.5))
curve(dnorm,from= -4,to=4,add=TRUE,col="red") # la loi N(0,1)

x<-seq(-5,5,0.1)
lines(x,dnorm(x,0,sqrt(2)),col="green4") #la lere loi propos?e
```

Histogram of z

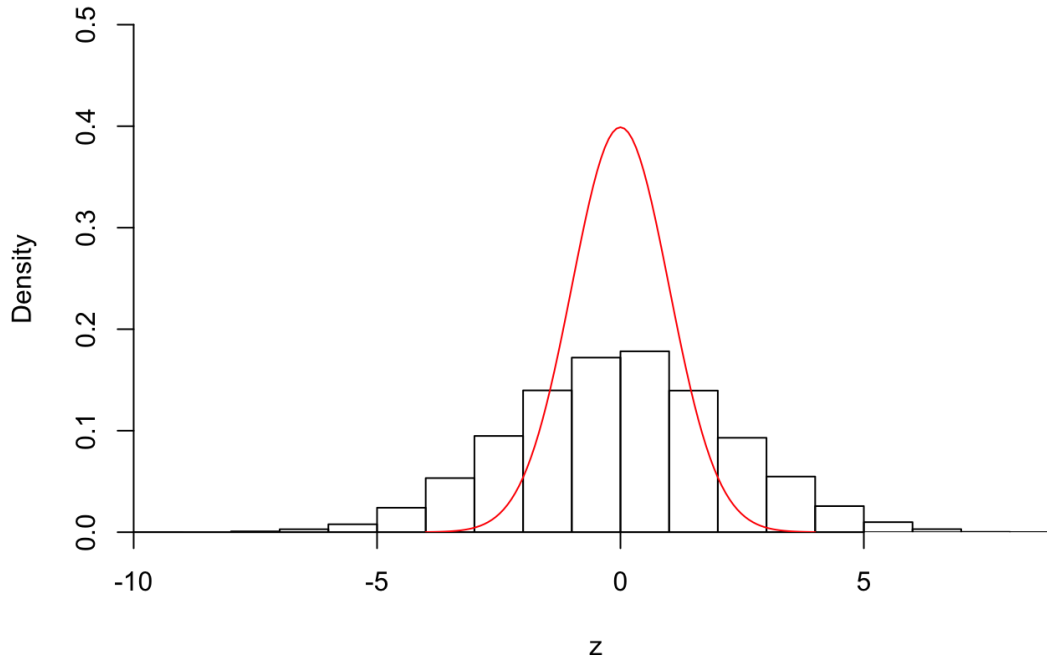


```
x<-seq(-5,5,0.1)
# lines(x,dnorm(x,##### a vous de trouver#####),col="blue4") # la loi que vous proposez (pensez à d
écommentez la ligne...)

# Deux variables gaussiennes de meme esperance mais pas de meme ecart type
x<-rnorm(10000,0,1)
y<-rnorm(10000,0,2)

z<-x+y
hist(z,breaks=20,freq=FALSE,ylim=c(0,0.5))
curve(dnorm,from= -4,to=4,add=TRUE,col="red")
```

Histogram of z



```
x<-seq(-5,5,0.1)
#lines(x,dnorm(x,##### a vous de trouver#####),col="green4") #la loi que vous proposez
```

$$X + Y \rightsquigarrow$$

$$X - Y \rightsquigarrow$$

Déduire de ce qui précède la loi de la variable $\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \dots + X_n)$ où toutes les variables X_i sont gaussiennes centrées réduites. Écrire un script pour l'illustrer.

```
#####
# Ecrire un script pour expliquer la loi d'une moyenne de lois gaussiennes #
#####
```

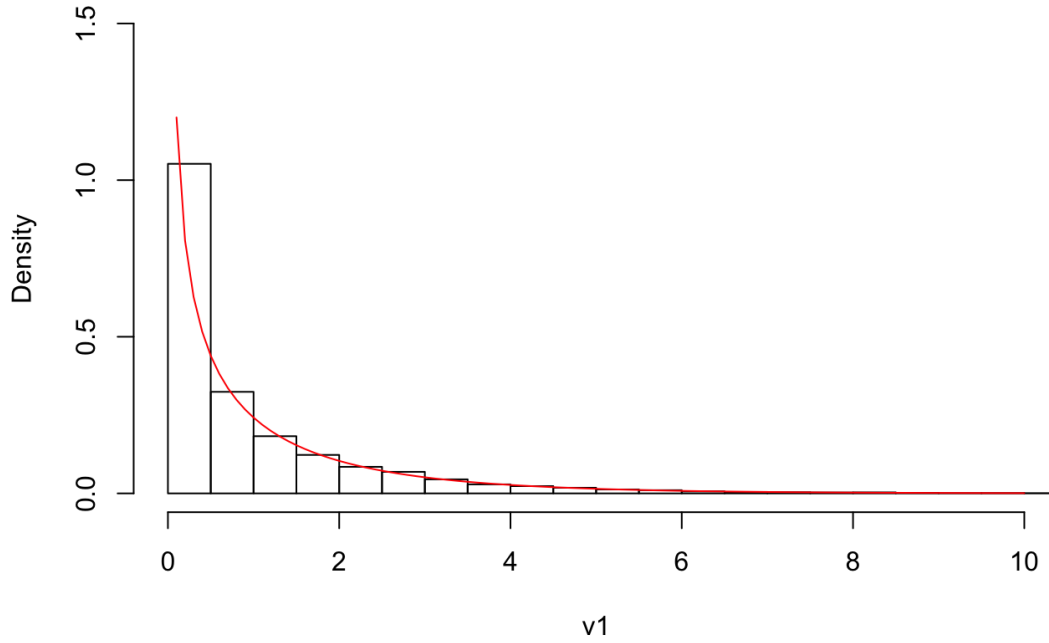
Somme de carrés de variables gaussiennes centrées réduites

Pour une somme de carrés de variables gaussiennes centrées réduites regarder la suite du script. Trouver le nom de la nouvelle loi des carrés et comment on choisit ses paramètres. (consulter l'aide de R)

```
#####
## Somme des carrés de loi gaussiennes #####
#####
v1<-rnorm(10000,0,1)^2
v2<-rnorm(10000,0,1)^2+rnorm(10000,0,1)^2
v3<-rnorm(10000,0,1)^2+rnorm(10000,0,1)^2+rnorm(10000,0,1)^2
v4<-rnorm(10000,0,1)^2+rnorm(10000,0,1)^2+rnorm(10000,0,1)^2+rnorm(10000,0,1)^2

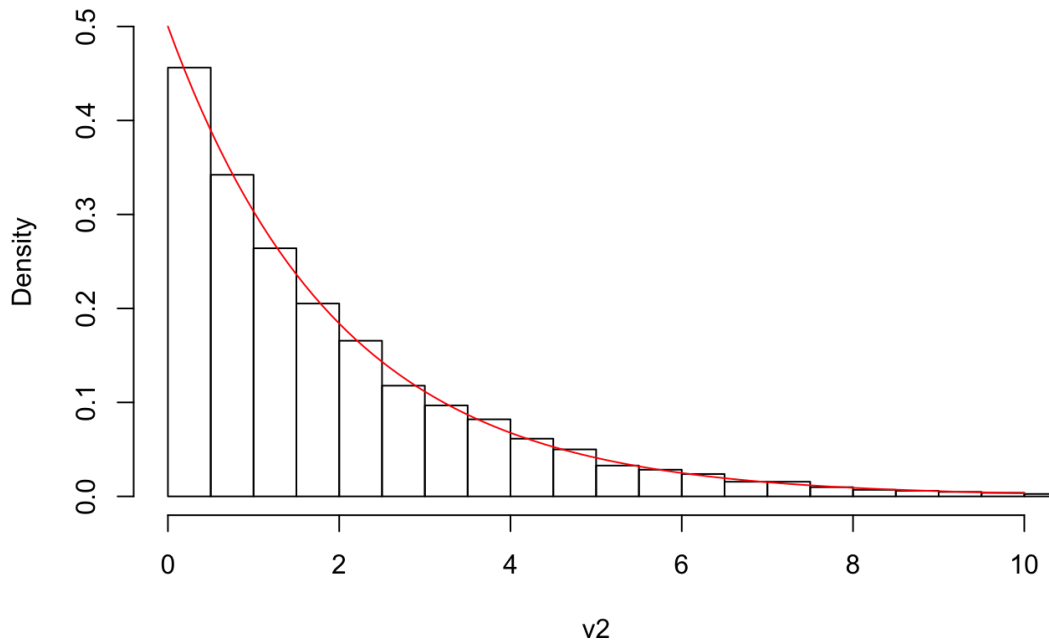
hist(v1,breaks=40,prob=TRUE,xlim=c(0,10),ylim=c(0,1.5))
curve(dchisq(x,1),from= 0,to= 10,add=TRUE,col="red")
```

Histogram of v1



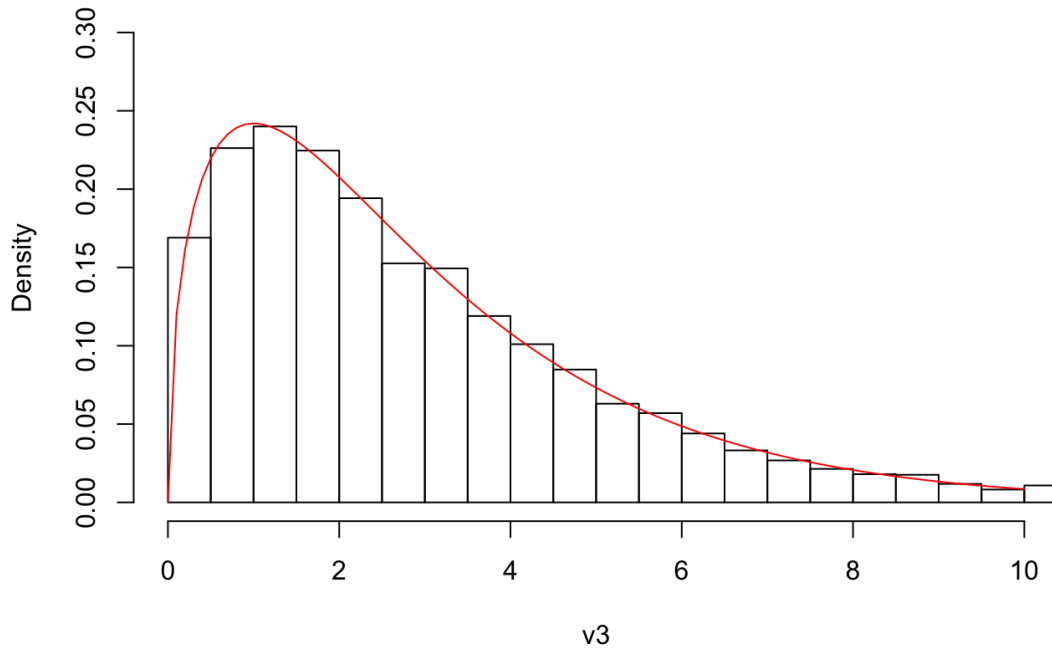
```
hist(v2,breaks=40,prob=TRUE,xlim=c(0,10),ylim=c(0,0.5))  
curve(dchisq(x,2),from= 0,to= 10,add=TRUE,col="red")
```

Histogram of v2



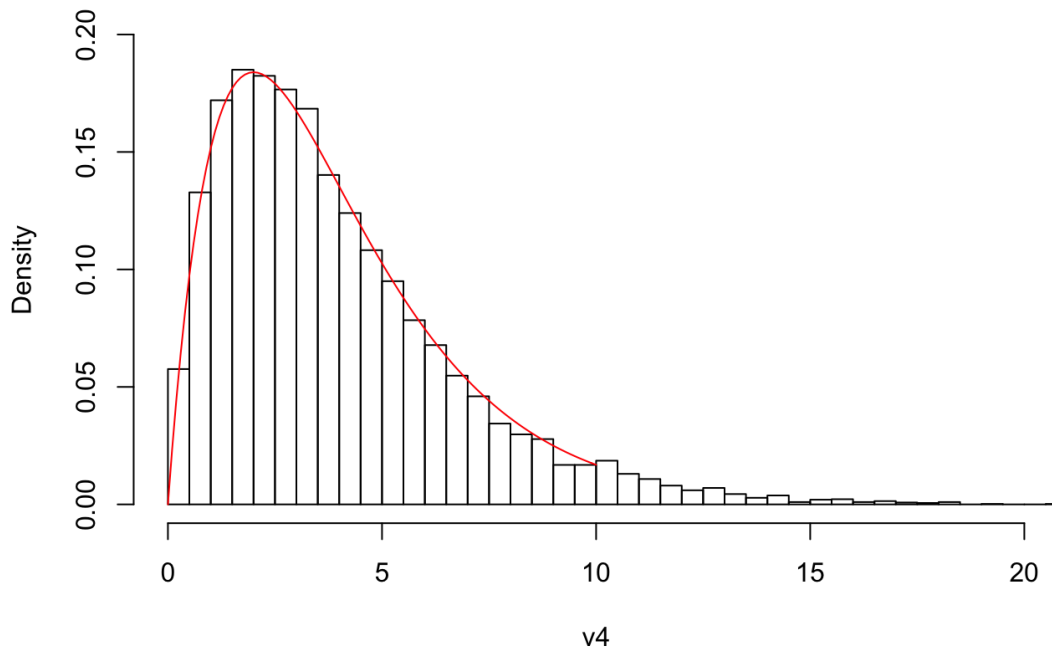
```
hist(v3,breaks=40,prob=TRUE,xlim=c(0,10),ylim=c(0,0.3))  
curve(dchisq(x,3),from= 0,to= 10,add=TRUE,col="red")
```

Histogram of v3



```
hist(v4,breaks=40,prob=TRUE,xlim=c(0,20),ylim=c(0,0.20))
curve(dchisq(x,4),from= 0,to= 10,add=TRUE,col="red")
```

Histogram of v4



Question : Etes-vous en mesure de compléter la formule ci-dessous ?

$$X_1^2 + X_2^2 + X_3^2 + \dots + X_n^2 \rightsquigarrow$$

Exercice

La taille X d'un homme français tiré au sort dans la population est supposée suivre une loi normale de moyenne 172 cm et de variance 196 cm^2 :

1. quelle est la probabilité que X soit inférieure à 160 cm ?

2. quelle est la probabilité qu'un homme tiré au sort mesure plus de deux mètres ?
3. Que vaut $P(165 < X < 185)$?
4. La taille des femmes françaises est modélisée par une loi gaussienne de moyenne 162 cm et de variance 144 cm^2 . Quelle est la probabilité qu'un homme tiré au sort soit plus grand qu'une femme choisie au hasard ?

Loi Exponentielle

On dit que X suit une loi exponentielle de paramètre $\lambda > 0$, ce que l'on note $X \rightsquigarrow \mathcal{E}(\lambda)$ ssi sa fonction densité est :

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

Question : Vérifier que f est bien une fonction densité.

La fonction de répartition de X est

$$\forall x \in \mathbb{R}, F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 1 - \exp(-\lambda x) & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

L'espérance d'une loi exponentielle est

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx = \frac{1}{\lambda}$$

Le moment d'ordre 2 d'une loi exponentielle est

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \frac{2}{\lambda^2}$$

et donc la variance vaut

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{\lambda^2}$$

On peut mettre en oeuvre cela en R

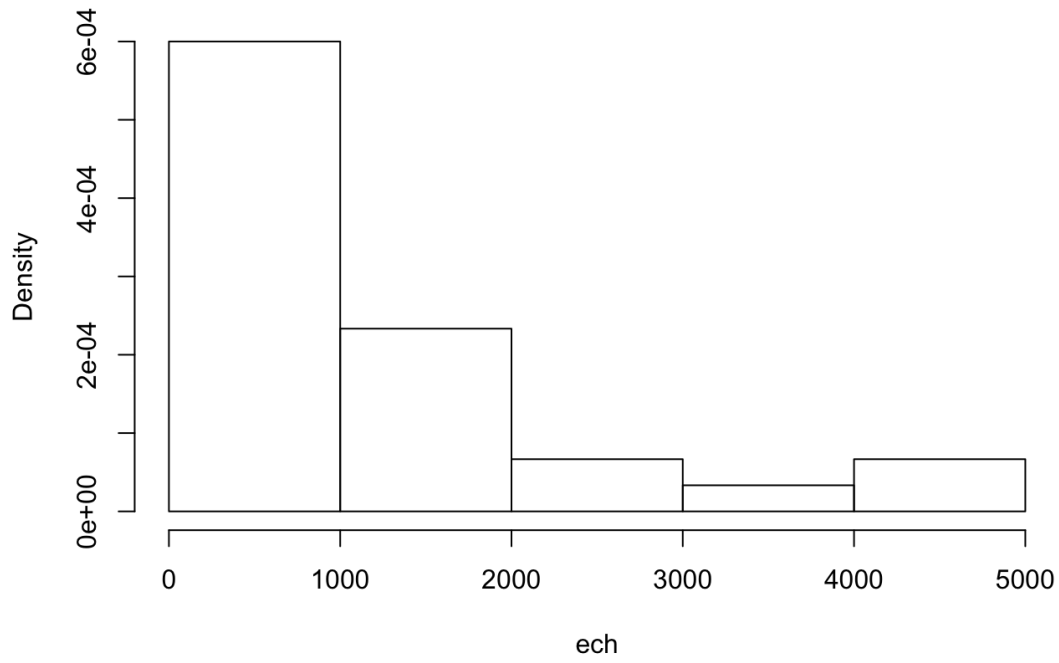
```
n <- 30           # taille de l'échantillon
lambda <- 10^{-3} # paramètre de la loi exponentielle
ech <- rexp(n, rate = lambda)
mean(ech)
```

```
## [1] 1182.928
```

```
var(ech)
```

```
## [1] 1543949
```

```
hist(ech, freq=F)
```

Histogram of ech

```
plot(ecdf(ech))
```

ecdf(ech)